Assignment 4 - CSC/DSC 462 - Fall 2018 - Due November 27

All questions of this assignment will make use of the `Cars93` data set from the `MASS` library. This library is included in the standard `R` distribution. To access and view this data set, use the commands:

```
> library(MASS)
> class(Cars93)
> head(Cars93)
> help(Cars93)
```

This data set contains various specifications of 93 car models. We will be interested in two of these: `MPG.city` and `Passengers`.

Q1: We first consider various transformation methods.

(a) Let $Y = (Y_1, \ldots, Y_n)$ be the observations in `MPG.city`. Create a histogram of Y, and comment on the skewness of the distribution.

(b) Create a function which implements the empirical rule. This function should input a single vector containing a sample, and output a 3×2 matrix summary. Rows 1 to 3 correspond to $k = 1, 2, 3$ standard deviations. The first column should contain the proportion of the data within k standard deviations of the mean, and the second column should contain the theoretical proportion for a normal distribution. The rows and columns should be suitably labelled. Test your function using a simulated random sample from a normal, exponential and uniform distribution. Use $n = 10000$ for each. Note that the choice of parameters will not make a difference, so just use the default values.

(c) Assuming right-skewness was detected in Y, we might consider a log-transformation of Y. However, the properties of the log-transformation (in particular, its ability to induce symmetry) can be affected by any data offset. So we might, more generally, consider the transformation $\log(Y_i + a)$ where a is some constant. The log-transformation can then be standardized by, for example, selecting $a = -\min(Y) + 1$, so that the lower bound of the transformed data $\log(Y_i - \min(Y) + 1)$ will be zero. To investigate this, consider the original data Y, and two transformations:

$$\begin{aligned} Y' &= \log(Y) \\ Y'' &= \log\left(Y - \min(Y) + 1\right) \end{aligned}$$

On a single plot window, plot a histogram and normal quantile plot for Y, Y' and Y'' (use a 3×2 plot grid, using one row for each version of the data). Then apply your empirical rule function to each version of the data. Is at least one version of the transformation able to induce an approximate normal distribution?

Q2: We consider the possibility that lack of normality can be caused by pooling heterogeneous sources of data.

(a) To prepare the data, use the `subset()` function to create a new data frame containing only cars with passenger capacity `Passengers` equal to 4,5,6 or 7. Create side-by-side boxplots of MPG rating Y, grouped by `Passengers`, and include in the same plot a single boxplot for the complete data Y. Comment on what you see.

(b) We might expect `MPG.city` to be negatively correlated with passenger capacity (that is, larger cars tend to have lower MPG ratings). So, we can normalize Y using means and standard deviations which are allowed to depend on the variables `Passengers`. Let $\bar{X}_m$ and S_m^2 be the sample mean and variance of Y restricted to the condition `Passengers` $= m$, $m = 4, 5, 6, 7$. Then for each observation Y_i create an adjusted Z-score

$$Z_i = \frac{Y_i - \bar{X}_m}{S_m},$$

where m is the value of `Passengers` for observation Y_i. Then, following Part (d) of **Q1**, plot a histogram and normal quantile plot for the transformed values $Z = (Z_1, \ldots, Z_n)$. Also, apply your empirical rule functions. Are the values Z approximately normal?

Q3: This question considers how to construct and interpret confidence intervals.

(a) Create a function which constructs a level $1-\alpha$ confidence interval for a mean μ given a random sample from a normal $N(\mu, \sigma^2)$ distribution. The function should accept a vector containing the data and α. Note that σ^2 should be assumed unknown.

(b) Apply the function of Part (a) to MPG observations Y separately for each value of `Passengers` $= m$, for $m = 4, 5, 6, 7$. Use confidence level 95%. Do the confidence intervals for `Passengers` $= 4$ and 5 overlap (that is, are there values of μ contained in both)?

(c) It can be shown that if a level $1-\alpha$ confidence interval for a mean doesn't contain 0, then a two-sided hypothesis test for null hypothesis $H_o : \mu = 0$ would reject H_o with an observed significance smaller than α. However, the situation for a difference in means is more complicated. Suppose we are given standard confidence intervals

$$\begin{aligned} \bar{X}_1 &\pm t_{n_1-1,\alpha/2}\frac{S_1}{n_1} \\ \bar{X}_2 &\pm t_{n_2-1,\alpha/2}\frac{S_2}{n_2}, \end{aligned}$$

for means μ_1, μ_2 respectively. Assume the samples used for each confidence interval are independent of each other. Then suppose the lower bound of the first confidence interval is larger than the upper bound of the second, so that the two do not overlap. Is this equivalent to rejecting the null hypothesis $H_o : \mu_1 = \mu_2$ based on the conventional T-statistic:

$$T = \frac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

assuming unequal variances? Why or why not?

(d) Do a formal two-sided T-test against null hypothesis $H_o : \mu_4 = \mu_5$, where μ_m is the mean MPG for cars with passenger capacity m. Assume that variances are not equal. Can we reject H_o with a $\alpha = 0.05$ significance level? Does this contradict the results of Part (b)?

Q4: While it seems clear that cars with larger passenger capacity have lower MPG, we may also consider fuel efficiency on a per-passenger basis. A 6 passenger car might have a rating of 19.3 MPG. Thus, one mile of travel consumes 1/19.3 gallons. However, if we assume for the moment that the vehicle is used to full passenger capacity, then we can say that each *passenger* consumes $1/(6 \times 19.3)$ gallons. We can therefore define a *passenger-mile per gallon* rating as

$$PMPG = N_P \times MPG,$$

where N_P is the number of passengers.

(a) Suppose we are given the following parameters and summary statistics for MPG ratings for $m = 4$ and $m = 5$ passenger cars:

	MPG	
	4 Passengers	5 Passengers
Population mean	μ_4	μ_5
Population variance	σ_4^2	σ_5^2
Sample size	n_4	n_5
Sample mean	$\bar{X}_4$	$\bar{X}_5$
Sample variance	S_4^2	S_5^2

Suppose we wish to construct a two-sided hypothesis test with a null hypothesis that the PMPG rating is the same for each class of car. Give the null and alternative hypotheses, as well as the appropriate T-statistic (assuming unequal variances). Use the quantities given in the preceding table.

(b) Carry out the T-test of Part (a) of this question. You can do this by creating the PMPG ratings directly as the product of the MPG ratings and the passenger capacities, then applying the `t.test()` function. Use an $\alpha = 0.05$ significance level, and assume unequal variances. How does your result compare to Part (d) of Question 3?

(c) Create side-by-side boxplots of the PMPG ratings, grouped by `Passengers`. As in Part (b) of Question 3, create level 95% confidence intervals for PMPG separately for each value of `Passengers` $= m$, $m = 4, 5, 6, 7$. Superimpose these directly on the boxplots. Does there seem to be a big difference in PMPG between different car classes? (**HINT:** The location of the ith boxplot on the horizontal axis is i.)